Date: June 21, 2006

To: Jim Anderson, DEQ NWR, Portland Harbor Project Manager

From: Jennifer Peterson, DEQ NWR, Portland Harbor Section, Toxicologist

RE: Portland Harbor Superfund Site Ecological Risk Assessment: Interpretive Report: Estimating Risks to Benthic Organisms Using Predictive Models Based on Sediment Toxicity Tests

General Comments:

Formatted: Font: Not Bold

1. **Inclusion of Appropriate Data in the Model:** The biggest comment relative to this report is the selection of appropriate data to use in the model. Early on in discussions regarding development of a predictive model, EPA partners as well as LWG consultants agreed we did not want to include data in a model when bioavailability was an issue. An example of this is where high concentrations are detected in the sediments, but these concentrations are bound up in a less bioavailable fraction such as pencil pitch. We had concerns about including T-4 in the analysis for this reason. Including these samples in the analysis can greatly skew the model results, because effect is not correlated with bioavailable fractions in the sediment. Based on my review of the report, I have concerns about the inclusion of GASCO effect / concentration data. This site has the potential to contain many different bound PAH contamination including pencil pitch. However, these samples were still included in the model analysis. This results in the inclusion of "no hits" with very high concentrations of PAHs. Looking at the highest no-hit concentrations, the top 6 samples are all in the vicinity of the GASCO site. Examples include G-264, 1,708,600 ppb total PAHs, G-301, 1,250,500 ppb, and G178, 470,060 ppb. Conditions off GASCO are confounded by the mixture in sediments of these less-bioavailable fractions such as pencil pitch and weathered tar pieces tar along with more fresh PAH and coal tar fractions that are more bioavailable and elicit effects. These two conditions may be teased out by a re-analysis of the sediment samples off the site. Conditions off GASCO can also lead to variance in the toxicity test results that are too high to detect anything but very large differences (low power), resulting in statistically indeterminate results. The GASCO site had the highest incidence of indeterminate samples at all effects levels (Figure 2-2). If these effects cannot be teased apart, we could simply omit samples off the GASCO site from the analysis. For this site, it may be clear that due to the variability in the forms of the contamination that we cannot accurately predict toxicity off this facility.

2. **Focus the Modeling Efforts:** This report recommends the focusing on the floating percentile method for future modeling efforts. I would

recommend focusing the modeling effort as well, but focusing on the logistic regression model. This will help minimize the review of the two models, and maintaining the appropriate optimization of the data that is specific to each model. Optimizing the floating percentile method may be particularly difficult. In contrast to the logistic regression model, the FPM does not use available data in order develop a quantitative relationship between concentration and response, but instead is a model heavily biased by initial conditions and properties of the dataset under evaluation.

3. **Data Gaps:** TPH was found to have good potential relationships with toxicity, but because we only analyzed TPH at a limited number of stations the model cannot assess this relationship (see page 21). This seems like a good candidate for collecting more information.

4. **Three Tiered Framework:** Based on the inherent reliability problems with attempting to come up with one SQV, DEQ would recommend the approach of calculating two screening values; a low screen below which a sample shouldn't be toxic and a high screen above which it should be toxic. We should be able to optimize the SQGs appropriately at these two ends of the spectrum, and are generally in agreement in predicting very toxic samples and those that are clearly non-toxic samples. However, we don't agree on the classification of the samples that fall in between these classifications. The values that fall in between these two classifications would be classified as "indeterminate", and would require empirical toxicity testing or the use of additional lines of evidence. The LRM is well suited to this. It could also be done with the FPM (as was done for the DRAFT Washington Freshwater criteria).

5. **Alternative Approaches For Subsets of PH Sediments:** As stated in the March 18th work plan (Section 9.2), there are areas for which the predictive approach would not apply in Portland Harbor. This could included the physical form of the contaminant (as mentioned above), or the localized presence of contaminations over smaller spatial scales in the ISA (e.g. pesticides around RM 7). The work plan states models or other approaches would be developed for these areas. However, this was not included in the report. Also, areas where volatile chemicals were detected in sediments and may be contributing to toxicity, but not evaluated in this report should be examined.

Specific Comments:

Page 1, Section 1.0, Introduction: There are statements made here that state that the sediment toxicity testing and derivation of sediment quality values (SQGs) form the primary lines of evidence for the benthic community, and that other lines of evidence such as tissue residue concentrations and comparison to

surface water and transition zone water concentrations would be secondary lines of evidence. This text should be revised, as the weights of different lines of evidence will be developed through the development of the weighting matrix.

Pages 5-6, Section 2.1.2, Biological Effects Definitions: There have been questions in past comment letters expressing concern about the selected alpha level for determining statistical significance. According to the work plan proceeding this report (*Estimating Risks to Benthic Organisms Using Sediment Toxicity Tests*, FINAL, dated March 18, 2005), an alpha level of 0.1 was to be used where it is found that test power of the dataset is low, according to ASTM guidelines (2003). Only an alpha level of 0.05 was used here. Since power is directly related to variance in the sample data, the variance in the analysis should be clearly reported and understood. To address concerns about the appropriateness of the statistical analysis to determine hits and no hits, it is recommended that the methodology outlined by Thursby et al., 1997 and Phillips et al., 2001 be followed. This approach more directly deals with issues that hinder appropriate statistical comparisons to determine statistical difference. This protocol considers performance over a large number of comparisons. MSD values are calculated to determine a critical threshold for statistically significant sample toxicity. Significant toxicity threshold values (as a percentage of laboratory control values) are presented for each species and endpoint based on the data.

Data should be reported as indicated in Table 2 of Phillips et al, 2001, which clearly shows the sample and control response, the sample response as a % of the control, MSD threshold, significance of t-test, and whether it was identified as toxic, non-toxic or indeterminate. This will improve the transparency of the statistical analysis, and will address several concerns associated with interpreting toxicity test data. These include:

1) The identification of small differences that are statistical different from the control, which may increase the probability of making a type I error (identifying a sample as toxic when in reality it is not). This reporting should eliminate cases where statistical significance is assigned in individual cases because the among-replicate variability is small in a more transparent fashion. It will allow for a better understanding for where and how much this occurred.

2) Samples with large variance in the data (e.g. variance lies outside the 10^{th} and 90^{th} percentiles) should be reported. Declaring a sample non-toxic in this case would lead to a greater probability of making a Type II error (saying it is non-toxic when in reality toxicity exists). This will help in understanding where areas of large variance occurred (e.g. where differences from the control exceed 20 to 25%), and further action in those areas such as re-testing.

3) This method more accurately describes Beta error through a graphic representation of the statistical power (1-B). For example, power curves can be developed using the 10th, 25th, 50th, 75th, and 90th percentiles of the variance. Power curves can be superimposed with curves showing the probability of statistical difference created from the cumulative frequency of calculated MSDs.
This approach should explicitly define MSD for the project in a non-arbitrary manner. A better understanding of the power curves relative to the data's variance aids in decisions regarding what difference from control is appropriate for determining statistical significance. Once a threshold for significance is determine, all of the test's data is included in the acceptability analysis for that test.

Page 6, Section 2.1.2, Statistical Difference Determinations: What analysis was used to determine statistical significance? The footnotes on Table 2-1 state the means of untransformed mortality or weight data was used in the definitions of effect levels. Were test and reference stations tested for normality? Were t-tests used?

Page 6, Section 2.1.2, Indeterminate Stations: We understand there may be situations where low power is a problem because the variance may be too high in the test replicates to detect anything but vary large differences. Since the test responses were compared to control responses for the statistical evaluation, it is likely large variability in response came from the test sediment. The source of the variance should be reported here, because it could be do to variations in bioavailability related to chemical form in the environment, or due to poor sediment homogenization prior to testing.

Page 6, Section 2.1.2, Biological Effects Definitions, Statistical Difference from Negative Control: For the floating percentile analysis, it would be still important to include a Level I effects level based on a statistical difference from negative control. Again, this may be more important for the floating percentile analysis (and AET derivation), especially since it is so reliant on the how we define no-hits, as apposed to hits (see page 7, second paragraph). Very small magnitude differences at the low end of the effects range may be very important for the development of the floating percentile model. The logistic regression model is not as sensitive to the omission of hits at the low range because it is the prevalence of toxic samples that primarily drive the curves, and the development of model relationships are not adversely affected by low power samples (Jay, correct me if I am wrong).

Formatted: Font: Not Bold

Page 7, Section 2.1.3, Use of Historical Toxicity Data: The objectives of the modeling effort are not just to improve model reliability as defined in the footnote on page 6 (correct predictions / total stations). The results of combining historical or regional data should be presented in how it changes the endpoints the

Formatted: Font: Not Bold

Formatted: Font: Not Bold

government team are interested in optimizing; including % Predicted No Hit Efficiency.

Page 8, Section 2.2.1, Data Quality: The text states that results with qualifier definitions listed in Table 2-3 were excluded. It looks like excluding samples with the "N" qualifiers excluded a lot of data (esp. pesticides). It should be confirmed that all PCB / DDT interferences in this dataset were properly re-analyzed according to previous EPA direction and the memo entitled "*EPA Region 10 Guidance for Data Deliverables from Laboratories Utilizing SW-846 Methods 8081 and 8082 from the Analyses of Pesticides and PCB Aroclors*". High detection limits, or elimination of "N" qualifiers that may represent interference problems can have a significant affect on the appropriateness of any model that attempts to correlate effects with sediment concentrations. This is particularly worrisome because the text states that this exclusion primarily affected the pesticide data, and that "between 35 and 67% of the summed data of DDD, DDE, DDT, total DDT, total chlordane, and total endosulfan were excluded" (in addition to between 23 to 53% aldrin, hexachlorocyclohexane, nonachlor, dieldrin, and methoxychlor).

Since the "N" qualifier is not an undetected value, it is unclear if this was an appropriate exclusion. It is also unclear why "N" qualifiers combined with "T" values were excluded. Is this the result of combining the results of two different analyses on one sample, where both of them were an "N"? Or, was one sample an "N" and the other a "J"? J values certainly shouldn't be excluded, so if they were combined with an "N" as a result of another analysis method shouldn't the "J" estimate take priority? In my mind, the N, NJ, and NJT values should be kept for modeling purposes.

Page 9, Section 2.2.2, Data Organization: For the FPM, aluminum and selenium should be added back into the model. The analysis shows that there is an association between aluminum and effects. I also wouldn't say that just because they are crustal elements that they are non-toxic, or that they cannot also be elevated anthropogenically. The ANOVA results for these chemicals need to be included in Table 5-2. If they are not associated with toxicity, they will drop out if appropriate.

Page 11, Section 2.3.3 and Table 2-4: For some chemicals where there were elevated detection limits, the exclusion of these chemicals in contributing to the sum could underestimate the total concentration. Was this examined?
A comment is made here that past comparisons have shown that using summations of certain compounds generates better results. It would be nice to have a citation listed there.

Page 12, Section 2.2.4, Normalization: It is unclear if normalization was tried for this dataset to determine if it improved the reliability estimates.

Page 14, Section 3.2, Reliability Analysis Results: Where are the results for the quotient methodology? Recent work has shown that quotients are a reliable way to develop SQGs. This is also an important methodology because it evaluates the cumulative toxicity of individual contaminants. This analysis was also specified in the work plan.

Page 18, Section 4.1, Last Paragraph: Why was magnitude of toxicity only evaluated for these chemicals that had greater then 50% detection frequency?

Page 24, Section 5.1, Floating Percentile Model: Step 4 mentions an issue that DEQ has asked about before. The analysis presented here, consistent with earlier FPM applications, splits the hit and no-hit groupings into different distributions before calculating the percentages. This does not seem like a necessary or justifiable approach. In applying the floating percentile method, it makes sense to include the larger entire dataset. Limiting the distribution to the smaller no-hit dataset would appear to reduce the ability of the method to refine the resulting optimized concentrations. DEQ instead recommends that a single distribution of all the data be used in calculating the percentages. The use of one distribution to develop screening values does change the SQVs and the reliability estimates. The false negative and predicted no-hit efficiency values are similar to the LWG results, but in some cases the false positive and predicted hit efficiency values are substantially different (higher). These concerns were raised early in the evaluation of this model with the LWG, but the report fails to comment or acknowledge these concerns.

In addition, how were the results for each station assigned a hit/no hit status to create the distributions? Was this based on statistical difference only?

Step 5 states that analytes were only retained for model development for each endpoint if they were associated with toxicity at two or three of the effects levels. Why was this limitation placed on the analysis? It seems like we would want to pursue associations of toxicity even if it was only at one effects level. For example, this criteria excludes the development of models for cadmium, diesel-range hydrocarbons, mercury, pentachlorophenol and total PAHs for the *Hyalella* growth endpoint.

If the distributions were determined not be statistically different, then the contaminants were assigned AET values. It seems like they may not be statistically different because of the variance in response, but it may still not be appropriate to select the highest no effect concentration as an AET. If this is due to variance, it may be more appropriate to rely on empirical tests to determine toxicity at a given location.

Step 6: All hand optimizations need to be documented.

Formatted: Font: Bold

Formatted: Font: Not Bold

Formatted: Font: Italic

Formatted: Font: Italic

Formatted: Font: Not Bold

Page 30-31, Section 5.1.1: In addition to using different distributions, as mentioned previously, there are a couple of other things that the DEQ FPM model handles differently. We have not had time to make the DEQ version of the model the same, but these could also be contributing to some of the differences that we have been seeing.

- When increasing the concentrations in an attempt to lower the number of false positives, we took the following steps : (1) find the chemical with the highest number of FPs; in case of a tie, use the chemical with the lower concentration in that step, (2) increase the concentration of that one chemical by the designated increment, (3) recalculate the %FN and #FP and (4) if %FN goes up, go back to previous step and consider that chemical completed, otherwise start over with step (1) until #FP reaches 0. From the description in this report, it appears that the LWG goes back to step (2) instead of (1). In other words, once the chemical with the highest #FP is selected, they keep raising the increments and recalculating the %FN and #FP until that can no longer be done for that chemical. At that point, they select the next highest #FP. This difference could result in our getting different values.

- **Page 31, 5th Bullet:** When increasing the %FN to the next level, they build on the values determined in the previous step. In the DEQ model, each step is independent. After 5% FN, it just starts over at 10% FN without any regard to the previous answers. DEQ was definitely not aware of this difference before submittal of this report, and this explains how they avoid getting answers that sometimes go up and down instead of going up or holding steady when[?] the %FN is increased. In previous versions of the FPM, there was up and down variability in the answers. This is a new step to prevent it from occurring with this dataset. However, artificially determining to build on values from the previous step seems arbitrary and may hide problem areas with the methodology.

- DEQ's main concern with the differences in our results is that fact that we do not get the same results for our performance measures and our results are not as high as reported here. This appears to be related to the fact that DEQ ends up with higher numbers of FPs, thus creating lower values for Efficiency (No-Hit Reliability) and Predicted Hit Reliability. This may also be tied in with the slight differences in steps mentioned above.

Page 43, Section 5.3, Logistic Regression Analysis: DEQ focused on the floating percentile methodology for this review. However, we did have a few broad comments on the approach, as listed below:

Formatted: Font: Bold

Formatted: Font: Not Bold

It appears that the LWG did not conduct a separate optimization (curve fitting) of the Pr_Max values with proportion of toxic samples. Jay Field should be able to clarify the appropriateness of the method used by the LWG. If a step was omitted, that would reduce the reliability of the results.

Pr_Prod has been dropped in previous evaluations as a reliable measure, but we would like to see an evaluation in LWG's report. On basic principles, we would expect to see this work well. Perhaps it will end up working better with the Portland Harbor dataset.

Page 43, Section 5.3.1: The "screened data set" method mentioned in Step 4 of the LRM Methodology should be evaluated to see if it would have any beneficial results for the FPM.

Formatted: Font: Not Bold

Formatted: Font: Bold